Combining Discriminant Models with new Multi-Class SVMs

Yann Guermeur

Combining Discriminant Models with new Multi-Class SVMs

Yann Guermeur, LORIA[1]

NeuroCOLT2 Technical Report Series

NC2-TR-2000-086

December, 2000

Produced as part of the ESPRIT Working Group
in Neural and Computational Learning II,
NeuroCOLT2 27150

[1]email: Yann.Guermeur@loria.fr LORIA INRIA-Lorraine, Campus Scientifique, B.P. 239, 54506 Vandœuvre-lès-Nancy cedex, France

Abstract

The idea of combining models instead of simply selecting the "best" one, in order to improve performance, is well known in statistics and has a long theoretical background. However, making full use of theoretical results is ordinarily subject to the satisfaction of strong hypotheses (weak correlation among the errors, availability of large training sets, possibility to rerun the training procedure an arbitrary number of times, etc.). In contrast, the practitioner who has to make a decision is frequently faced with the difficult problem of combining a given set of pretrained classifiers, with highly correlated errors, using only a small training sample. Overfitting is then the main risk, which cannot be overcome but with a strict complexity control of the combiner selected. This suggests that SVMs, which implement the SRM inductive principle, should be well suited for these difficult situations. Investigating this idea, we introduce a new family of multi-class SVMs and assess them as ensemble methods on a real-world problem. This task, protein secondary structure prediction, is an open problem in biocomputing for which model combination appears to be an issue of central importance. Experimental evidence highlights the gain in quality resulting from combining some of the most widely used prediction methods with our SVMs rather than with the ensemble methods traditionally used in the field. The gain is increased when the outputs of the combiners are post-processed with a simple DP algorithm.

Keywords: Classifier fusion, statistical learning theory, generalization performance, Support Vector Machines, protein secondary structure prediction, hierarchical sequence processing systems.

1 Introduction

Since the early sixties, and precisely the studies of Bates and Granger [6, 30], model combination has proved to be a viable alternative to model selection for a wide range of statistical inference problems. Theory in the field has made rapid strides [20, 21, 30, 56, 50, 45, 55, 43], however, till recently, theoretical evidence had been mainly developed in the framework of regression, whereas discrimination was seldom considered independently. In the last decade, many studies have dealt with the specific problems of discrimination, such as the estimation of Bayes error [73, 31], or variance reduction [71], the link between error correlation and error reduction [72] (see also [17]) as well as the decomposition of the error into a bias and a variance term [12]. The success of methods such as *bagging* [11] and *boosting* [65] has highlighted the usefulness of implementing bootstrap algorithms to improve the performance of "weak classifiers". This is of primary importance indeed, since the theory of boosting meets Vapnik's theory of bounds through the fundamental notion of *maximal margin classifier*. Classifier combination is thus currently endowed with a rich theoretical framework, which is very usefull as long as the problem at hand satisfies the hypotheses on which it is grounded. Unfortunately, in many real-life situations, the practitioner is faced with the worst configuration one can think of when combining models (pretrained experts with different types of outputs, errors highly correlated, small set of labelled data available for training, etc.). These difficulties prevent him from making the best of the potential of the theory, and his main concern is to avoid overfitting. As a consequence, in this context, the problem to be solved consists in finding a combiner of adequate complexity, so that, with high probability, the training error observed could constitute an estimate precise enough of the generalization error, and the gain in prediction accuracy, small as it should be, could be "guaranteed". This is precisely the type of situations for which Support Vector Machines (SVMs) have been developed. SVMs have been introduced by Vapnik and co-workers [18] as a direct implementation of the Structural Risk

Minimization (SRM) inductive principle. The aim of the support vector method, a description of which can be found for instance in [76, 15, 35, 19], is to maximize the generalization capabilities by minimizing an upper bound on the *expected risk* (or generalization error) with respect to the values of the model paramaters. This bound is systematically made up of two terms. The first one is the *empirical risk* (training error), the second one, that Vapnik calls a confidence interval, is a growing function of the *capacity* of the model, capacity which can be expressed in terms of different measures, the most common one being the Vapnik-Chervonenkis (VC) dimension [77]. Simple introductions to the theory of bounds applied to neural networks can be found in [37, 3]. With this structure of the bound in mind, it appears immediately that the SRM inductive principle can be implemented by minimizing the control term for different levels of the empirical risk, in order to find a minimum of the *guaranteed risk* functional with a linesearch. Indeed, this aim is reached with the support vector methods developed for estimating indicator or real-valued functions. Unfortunately, although many multi-class discriminant models have been developed around the support vector method, none of them owns this property. Initially, multi-class discrimination was implemented with SVMs through the so-called *one-againts-the-rest* or *one-per-class* approaches [66, 75]. Later on came the *pairwise-coupling* decomposition scheme [53, 78] and the k-class SVM proposed independently by Vapnik [76], Weston and Watkins [78], and Bredensteiner and Bennett [10], among others. Strictly speaking these three approaches fail to implement the SRM inductive principle, since they are not related, at least explicitly, to a uniform convergence result, or guaranteed risk, which makes it impossible to characterize a satisfactory compromise between training performance and complexity. In this article, building upon the uniform strong law of large numbers introduced in [24], we develop a theoretical framework which leads to the specification of a family of multi-class SVMs, each of which corresponds to a different expression of the guaranteed risk. This enables us by the way to provide Vapnik's k-class SVM with a theoretical grounding. Two of these SVMs are assessed as classifier combiners on an open real-world problem: the problem of protein secondary structure prediction. This task is of central importance in predictive structural biology. Numerous methods have been proposed to predict the secondary structure (see [22, 62] for reviews on the subject). *A priori*, implementing a combination of models appears particularly relevant in this context, since most of the prediction systems developed so far ordinarily use, in addition to the amino acid sequences, data from different knowledge sources (physicochemical properties, homology, etc.). Consequently, whenever secondary structure is to be predicted, several sets of conformational scores are available, which can be expected not to be utterly correlated. Indeed, most of the current best predictions methods already implement conformational score combination at one stage or another. This combination can take many forms, ranging from the simple linear opinion pool [63] to the more complex non-linear regression schemes performed by neural networks [82, 61]. Symbolic methods based on empirical results have also been implemented, such as the algorithm combine [7]. However, a constant of these studies is that the choice of a particular combiner is hardly ever justified, although it appears to have a crucial effect on performance. Furthermore, the scores combined are systematically homogeneous, i.e. they represent estimates of the same quantities, whereas the practitioner who needs to make his own prediction based on the results of several methods has most often to deal with inhomogeneous scores. Last but not least, the gain resulting from the combination is seldom significantly superior to the one resulting from a simple averaging of the outputs of the base classifiers. This phenomenon is indeed acknowledged by leading experts in the domain (B. Rost and G. Pollastri, personnal communications). A first attempt to overcome these limitations was described in [34]. In this paper, we establish that noticeable benefits can spring from combining protein secondary structure

models with multi-class SVMs. The gain in prediction accuracy over other standard ensemble methods becomes statistically significant with confidence exceeding 0.98 when the outputs are post-processed with a simple Dynamic Programming (DP) algorithm borrowed from the field of speech processing, which suggests that our SVMs would perform best when incorporated in hierarchical prediction systems. The organization of this paper is as follows. In section 2, we briefly summarize our uniform convergence result and explain the way it can be of practical use to study the generalization capabilities of multi-class discriminant models (bound the expected risk). In Section 3, these formula are applied to the multivariate linear regression model, which leads to the specification of the new multi-class SVMs. Initial experimental results, regarding the sole combination, are given in Section 4. The comparative study is developed in Section 5, where the possibility of post-processing the outputs is assessed.

2 Guaranteed Risk for Multi-class Discriminant Models

2.1 Framework of the study

We consider the case of a Q-category pattern recognition problem. Let $\mathcal{X}$ be the space of description (or input space) and $\mathcal{C}$ the set of categories. We make the assumption, standard in statistical learning theory, that there is a joint probability, fixed but unknown, on $\mathcal{X} \times \mathcal{C}$. Our goal is then to find, in the set $\mathcal{H} = \{h\}$ of functions implemented by a statistical model, a function which corresponds to the lowest error rate. The decision function associated with this function must thus be as close as possible to Bayes' decision rule. We make further the hypothesis that the elements of $\mathcal{H}$ are multivariate real-valued functions. Precisely, for each example x in $\mathcal{X}$ and each category C_k in $\mathcal{C}$, $(1 \leq k \leq Q)$, a function h_k of x taking its values in $\mathbb{R}$ is computed. The discriminant function associated with these regression functions is obtained by assigning each pattern x to the category C_k satisfying: $h_k(x) = \max_l h_l(x)$. This framework is indeed very common. In the case where the $h_k(x)$ are estimates of the class posterior probabilities, which occurs for instance when the model is a neural network and the training criterion is adequately selected [60, 8], choosing this decision function simply amounts to implementing Bayes' estimated decision rule. In what follows, $C(x_i)$ will denote indifferently the category of pattern x_i, or the index of this category, while y_i will be the corresponding canonical codings, i.e. $C(x_i) = C_l \Longleftrightarrow y_i = [y_{ik}] \in \mathbb{R}^Q$, where $y_{ik} = -1^{1-\delta_{kl}}$, and δ is Kronecker's symbol.

2.2 Uniform strong law of large numbers based on a covering number

In this context, we have established a uniform strong law of large numbers which is based of the following definitions.

Definition 1 (Covering number) *Let (E, ρ) be a pseudo-metric space, and $B(v, r)$ the closed ball of radius r and centre v in E. The* covering number *$\mathcal{N}(\epsilon, H, \rho)$ of a set $H \subset E$ is the smallest cardinality of the sets $\bar{H} \subset E$ such that*

$$H \subset \bigcup_{v \in \bar{H}} B(v, \epsilon)$$

The sets $\bar{H}$ satisfying this property are called ϵ-covers of H: each element in H is at a distance less than ϵ of an element in $\bar{H}$.

See [48, 16] for the fundamental results regarding covering numbers.

Definition 2 *Let $\mathcal{F}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For a set s of points in $\mathcal{X}$, define the pseudo-metric $d_{l_\infty,l_1(s)}$ on $\mathcal{F}$ as:*

$$\forall (f,\bar{f}) \in \mathcal{F}^2,\ d_{l_\infty,l_1(s)}(f,\bar{f}) = \max_{x\in s} \sum_{k=1}^{Q} \left| f_k(x) - \bar{f}_k(x) \right|$$

Definition 3 *For a given function h in $\mathcal{H}$, let $M_1(h,x)$ be the smallest index l such that $h_l(x) = \max_k h_k(x)$ and $M_2(h,x)$ the smallest index $l \neq M_1(h,x)$ such that $h_l(x) = \max_{k\neq M_1(x)} h_k(x)$. Define $\Delta h = [\Delta h_k]$, $(1 \leq k \leq Q)$, as the function from $\mathcal{X}$ into $\mathbb{R}^Q$, satisfying*

$$\Delta h_k(x) = \begin{cases} \frac{1}{2}\left(h_k(x) - h_{M_2(h,x)}(x)\right) & \text{if } k = M_1(h,x) \\ \frac{1}{2}\left(h_k(x) - h_{M_1(h,x)}(x)\right) & \text{otherwise} \end{cases}$$

Note that this function is directly related to the notion of margin introduced by Schapire and co-workers in [65], in order to extend to the multi-class case the uniform convergence results established for boosting algorithms. Define the threshold function $sign:\ \mathbb{R} \to \{-1,1\}$ as

$$sign(x) = \begin{cases} 1 & if\ x \geq 0 \\ -1 & otherwise \end{cases}$$

For $\gamma \in]0,1]$, define $\pi_\gamma : \mathbb{R} \to [-\gamma,\gamma]$ as the piecewise-linear squashing function

$$\pi_\gamma(x) = \begin{cases} \gamma.sign(x)\ if\ |x| \geq \gamma \\ x\ otherwise \end{cases}$$

$\forall h \in \mathcal{H},\ \Delta h^\gamma = [\Delta h_k^\gamma] = [\pi_\gamma \circ \Delta h_k]$, $(1 \leq k \leq Q)$. $\Delta\mathcal{H}^\gamma = \{\pi_\gamma(\Delta h)\ /\ h \in \mathcal{H}\}$. With these definitions at hand, we denote

Definition 4

$$\mathcal{N}_{\infty,1}(\gamma/2, \Delta\mathcal{H}^\gamma, 2N) = \max_{s_{2N}\in\mathcal{X}^{2N}} \mathcal{N}(\gamma/2, \Delta\mathcal{H}^\gamma, d_{l_\infty,l_1(s_{2N})})$$

In order to choose the optimal function $h \in \mathcal{H}$, we make the assumption that a training set $s_N = \{(x_i,y_i)\}$, $(1 \leq i \leq N)$, made up of labelled examples, iid according to the joint distribution on $\mathcal{X}\times\mathcal{C}$ to be infered, is available. Extending a definition from Bartlett [5], we introduce the following definition.

Definition 5 *The empirical risk with margin $\gamma \in]0,1]$ on a training set s_N of size N is*

$$R_{s_N}^\gamma(h) = \frac{1}{N}\left|\left\{(x_i,y_i) \in s_N\ /\ \Delta h_{C(x_i)}(x_i) < \gamma\right\}\right|$$

Studies on the use of margins in statistical learning theory date back from the early works of Vapnik [74]. Different illustrations of the richness of this approach can be found for instance in [67, 68, 65, 24]. In this context, extending Lemma 4 and Corollary 9 from [5], as well as Theorem 4.1 from [76], we established in [24] the following theorem (Corollary 20):

Theorem 1 *With probability at least $1-\delta$, the risk $R(h)$ of a function h computed by a numerical Q-class discriminant models $\mathcal{H}$ trained on a set of size N is bounded above by:*

$$R(h) \leq R_{s_N}^{\gamma}(h) + \sqrt{\frac{1}{2N}\left(\ln(2\mathcal{N}_{\infty,1}(\gamma/2, \Delta\mathcal{H}^{\gamma}, 2N)) + \ln(\frac{2}{\gamma\delta})\right)} + \frac{1}{N} \qquad (1)$$

The proof mainly follows the sketch of the proof of the classical Glivenko-Cantelli theorem [58] (see [24] for details). Note that, contrary to other well known bounds, this theorem does not rest on the hypothesis that the functions in $\mathcal{H}$ take their values in $[-1,1]^Q$, what makes it quite general. It applies for instance to multi-layer perceptrons, even when they have linear output units, and consequently also applies to SVMs. The bound is significantly tighter than the one obtained by using as capacity measure multi-class extensions of the VC dimension such as the *graph dimension* or *Natarajan dimension* [54]. A preliminary comparative study on this question can be found in [32]. To implement the SRM inductive principle, the main term, except from the empirical risk with margin γ, is the covering number $\mathcal{N}_{\infty,1}(\gamma/2, \Delta\mathcal{H}^{\gamma}, 2N)$ with characterizes the model capacity. Expressing this measure in terms of the model parameters can thus provide us with the objective function of the optimization problem corresponding to the implementation of the SRM inductive principle.

2.3 Bound on the covering number

Several methods have been proposed to bound covering numbers (see for instance [38, 2, 5]). In this section, we derive an upper bound on the covering number for general multi-class models (unspecified families of functions $\mathcal{H}$ taking their values in $\mathbb{R}^Q$), using the method introduced in [79, 69]. A key feature of this approach is that it directly bounds the covering numbers of interest rather than making use of a combinatorial dimension such as the extensions of the VC dimension cited before or the *fat-shattering dimension* [47]. To that end, we make the additional assumption that $\mathcal{H}$ is included in a finite-dimensional Banach space $E_{\mathcal{H}}$. For a given set $s_{2N} \in \mathcal{X}^{2N}$, $E_{\mathcal{H}}$ is endowed with the pseudo-norm derived from $d_{l_\infty, l_1(s_{2N})}$ in the canonical way. We assume further that $\mathcal{H}$ is bounded in $E_{\mathcal{H}}$ for all these norms. This is indeed a mild hypothesis, since it is satisfied among others by SVMs, regularization networks [29] and linear models, when prior information is assumed or is given. As a consequence, $\mathcal{H}$ is *precompact* (see [16] for a proof of this proposition), which means that the covering number of interest will always be finite. For the set $s_{2N} \in \mathcal{X}^{2N}$ aforementioned, let us define the following linear operator:

$$\begin{array}{rccc} T_{s_{2N}}: & \mathcal{H} \subset E_{\mathcal{H}} & \longrightarrow & M_{2N, \frac{Q(Q-1)}{2}}(\mathbb{R}) \\ & h = [h_k] & \mapsto & T_{s_{2N}}(h) \end{array}$$

with

$$T_{s_{2N}}(h) = \begin{bmatrix} h_1(x_1) - h_2(x_1) & \dots & h_k(x_1) - h_l(x_1) & \dots & h_{Q-1}(x_1) - h_Q(x_1) \\ \dots & \dots & \dots & \dots & \dots \\ h_1(x_i) - h_2(x_i) & \dots & h_k(x_i) - h_l(x_i) & \dots & h_{Q-1}(x_i) - h_Q(x_i) \\ \dots & \dots & \dots & \dots & \dots \\ h_1(x_{2N}) - h_2(x_{2N}) & \dots & h_k(x_{2N}) - h_l(x_{2N}) & \dots & h_{Q-1}(x_{2N}) - h_Q(x_{2N}) \end{bmatrix}$$

where i is the index of line ($1 \leq i \leq 2N$) and the one-to-one mapping between any couple (k,l) satisfying ($1 \leq k < l \leq Q$) and the column index j is given by the

equation

$$j = (k-1)Q - \frac{k(k+1)}{2} + l$$

Let $B_{\mathcal{H}}$ be a closed ball of $E_{\mathcal{H}}$ in which $\mathcal{H}$ is included. We endow $M_{2N,\frac{Q(Q-1)}{2}}(\mathbb{R})$ with the following norm:

$$\forall A \in M_{2N,\frac{Q(Q-1)}{2}}(\mathbb{R}),\ A = [a_{ij}]\,,\ \|A\|_{l_\infty,l_1} = \max_{1\leq i\leq 2N} \sum_{j=1}^{\frac{Q(Q-1)}{2}} |a_{ij}|$$

After some algebra, it comes:

$$\mathcal{N}(\gamma/2, \Delta\mathcal{H}, d_{l_\infty,l_1(s_{2N})}) \leq \mathcal{N}(\gamma/2, T_{s_{2N}}(B_{\mathcal{H}}), \|.\|_{l_\infty,l_1})$$

Since π_γ satisfies the Lipschitz condition:

$$\forall (f,\bar{f}) \in \mathcal{F}^2,\ \forall s \subset \mathcal{X},\ d_{l_\infty,l_1(s)}\left(\Delta f^\gamma, \Delta \bar{f}^\gamma\right) \leq d_{l_\infty,l_1(s)}\left(\Delta f, \Delta \bar{f}\right)$$

it comes readily

$$\mathcal{N}(\gamma/2, \Delta\mathcal{H}^\gamma, d_{l_\infty,l_1(s_{2N})}) \leq \mathcal{N}(\gamma/2, \Delta\mathcal{H}, d_{l_\infty,l_1(s_{2N})})$$

and finally:

Theorem 2 *For all $s_{2N} \in \mathcal{X}^{2N}$ and for all $\gamma \in]0,1]$,*

$$\mathcal{N}(\gamma/2, \Delta\mathcal{H}^\gamma, d_{l_\infty,l_1(s_{2N})}) \leq \mathcal{N}(\gamma/2, T_{s_{2N}}(B_{\mathcal{H}}), \|.\|_{l_\infty,l_1}) \tag{2}$$

We have thus reduced the problem of bounding the covering number appearing in (1) to the problem of finding an upper bound of $\mathcal{N}(\gamma/2, T_{s_{2N}}(B_{\mathcal{H}}), \|.\|_{l_\infty,l_1})$, when s_{2N} describes the whole set $\mathcal{X}^{2N}$. This can be done readily thanks to functional analysis results. Let E be a Banach space and let $U_E = \{v \in E / \|v\|_E \leq 1\}$ be its unit ball. Applying proposition 1.3.1. in [16], we get:

Theorem 3 *If a linear operator $T \in L(E,F)$ acting between arbitrary real Banach spaces E and F is of rank r_T, then*

$$\mathcal{N}(\epsilon, T\left(U_E\right), \|.\|_E) \leq \left(\frac{4\|T\|_F}{\epsilon}\right)^{r_T} \tag{3}$$

where $\|T\|_F$ is the standard operator norm given by: $\|T\|_F = \sup_{f\in U_E} \|T(f)\|_F$.

Let $f_o \in E_{\mathcal{H}}$ and $r \in \mathbb{R}_+$ be such that $B_{\mathcal{H}} \subset B(f_o, r)$. Then, Theorem 3 can be used to bound $\mathcal{N}(\gamma/2, T_{s_{2N}}(B_{\mathcal{H}}), \|.\|_{l_\infty,l_1})$ due to the following equation:

$$T\left(B(f_o, r)\right) = rT\left(U_E\right) + T(f_o) \tag{4}$$

This means that, once one has been able to characterize a ball of $E_{\mathcal{H}}$ in which the functions of $\mathcal{H}$ live, the nature of its parameters (centre and radius) rises no (theoretical) difficulty to bound the covering number of interest (keeping in mind that the smaller the radius is, the better the bound will be). Once a bound on the covering number of the linear operator is available, suffice it to take its supremum over the whole set $\mathcal{X}^{2N}$ to obtain a bound on $\mathcal{N}_{\infty,1}(\gamma/2, \Delta\mathcal{H}^\gamma, 2N)$.

3 Multi-Class SVMs

The results presented in the previous section apply to any multi-class discriminant system obtained by combining a multivariate model with Bayes' estimated decision rule. In this section, we turn to the specific case of multi-class SVMs. The study of the standard (bi-class) SVMs is usually done in two steps: first, the linear case (optimal hyperplane), then the non-linear one (by introduction of kernels satisfying Mercer's conditions [1]). Indeed, the specification of the training procedure does not take into account explicitely the nature of the kernel, although bounds on the generalization error of kernel machines have been derived (see for instance [79] for a very powerful theoretical framework on the subject). In the same way as a "linear" SVM shares the architecture of the perceptron, a multi-class linear SVM is a multivariate linear regression model (a set of hyperplanes of cardinality equal to the number of classes). In order to apply the SRM inductive principle to multi-class SVMs, and consequently to determine the objective function of the training procedure, we must thus bound the covering number of the multivariate linear (affine) model.

3.1 Covering number of the multivariate linear regression model

We assume that the data live in $\mathbb{R}^d$ $\left(\mathcal{X} \subseteq \mathbb{R}^d\right)$. Let $\mathcal{H} = \{h\}$ be the set of functions computed by this model, with

$$h(x) = Wx + b = \begin{bmatrix} w_1^T \\ \vdots \\ w_k^T \\ \vdots \\ w_Q^T \end{bmatrix} x + \begin{bmatrix} b_1 \\ \vdots \\ b_k \\ \vdots \\ b_Q \end{bmatrix}$$

To implement the SRM inductive principle, we must express a bound on the capacity measure of interest in terms of the constraints set on the matrix W and the vector b. The model is affine. In order to achieve our goal, it appears simpler to adopt a strategy advocated in [79], consisting in considering first the case of the corresponding linear model $\tilde{\mathcal{H}}$ computing the set of functions $\left\{\tilde{h}\right\}$. The possibility to estimate one covering number from the other springs from the following theorem, which extends a result established in [79], Section 9.1.

Theorem 4 *Let $\tilde{\mathcal{F}}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$ and $\mathcal{F}$ a set of functions satisfying: $\forall f \in \mathcal{F}$, $\exists(\tilde{f}, b) \in \tilde{\mathcal{F}} \times [-B, B]^Q$ / $f = \tilde{f} + b$. Let $\Delta\mathcal{F}$ and $\Delta\tilde{\mathcal{F}}$ be the sets of functions derived from $\mathcal{F}$ and $\tilde{\mathcal{F}}$ respectively, by applying Definition 3. Then the following bounds hold:*

$$\mathcal{N}_{\infty,1}(Q\epsilon, \mathcal{F}, 2N) \leq \left(\left\lceil \frac{2B}{\epsilon} \right\rceil + 1\right)^Q \mathcal{N}_{\infty,1}(\epsilon, \tilde{\mathcal{F}}, 2N)$$

$$\mathcal{N}_{\infty,1}(Q\epsilon, \Delta\mathcal{F}, 2N) \leq \left(\left\lceil \frac{4B}{\epsilon} \right\rceil + 1\right)^Q \mathcal{N}_{\infty,1}(\epsilon, \Delta\tilde{\mathcal{F}}, 2N) \tag{5}$$

We have made the hypothesis that $\mathcal{H}$ was bounded in $E_{\mathcal{H}}$, which implies that $\mathcal{X}$ is bounded in $\mathbb{R}^d$. Let K_x be such that: $\forall x \in \mathcal{X}, \|x\|^2 \leq K_x$, where $\|.\|$ stands for the

euclidian norm.

$$\forall s_{2N} \in \mathcal{X}^{2N},\ \forall h \in \mathcal{H},\ \|T_{s_{2N}}(\tilde{h})\|_{l_\infty,l_1} = \max_{1\leq i\leq 2N} \sum_{k<l} \left|(w_k - w_l)^T x_i\right|$$

Applying Cauchy-Schwarz inequality, it comes

$$\|T_{s_{2N}}(\tilde{h})\|_{l_\infty,l_1} \leq K_x \sum_{k<l} \|w_k - w_l\|$$

and consequently

$$\|T_{s_{2N}}(\tilde{h})\|_{l_\infty,l_1} \leq K_x \sqrt{\frac{Q(Q-1)}{2}} \sqrt{\sum_{k<l} \|w_k - w_l\|^2} \tag{6}$$

To sum up, combining the results exposed in (1), (2), (3), (4), (5) and (6), it appears that the confidence interval which constitutes, with the empirical risk, the expression of the guaranteed risk, is an increasing function of $\sum_{k<l} \|w_k - w_l\|^2$. This result extends nicely Vapnik's well known bound on the VC dimension of canonical hyperplanes in terms of the square of the norm of the corresponding vector (see [76], Theorem 10.3).

3.2 Unification of the multi-class SVMs proposed so far

Making use of the main result of the preceding section, we can readily specify a multi-class SVM, the objective function of which takes the confidence interval into account through the term $\sum_{k<l} \|w_k - w_l\|^2$. The training procedure associated with this model consists in solving the following quadratic programming problem:

Problem 1

$$\min_{h\in\mathcal{H}} \sum_{k,l=1}^{Q} \|w_k - w_l\|^2 + \frac{C}{N} \sum_{i=1}^{N} \sum_{k=1}^{Q} \xi_{ik}$$

$$s.t. \begin{cases} (w_{C(x_i)} - w_k)^T x_i + b_{C(x_i)} - b_k \geq 1 - \xi_{ik}, & (1 \leq i \leq N), (1 \leq k \neq C(x_i) \leq Q) \\ \xi_{ik} \geq 0, & (1 \leq i \leq N), (1 \leq k \neq C(x_i) \leq Q) \end{cases}$$

As usual, the non-negative slack variables ξ_{ik} have been introduced to take into account the fact that the data could be non-separable by the multivariate linear model. Their values characterize the empirical risk. Many algorithms are available to find an optimal solution (see for instance [25, 69, 23]). In fact, additional specifications are required in order to ensure the unicity of the optimal solution, due to the following result. Let $\left(W^{(1)}, b^{(1)}\right)$ be an optimal solution of Problem 1. Then the couple $\left(W^{(2)}, b^{(2)}\right)$ such that $w_k^{(2)} = w_k^{(1)} + v$, $(1 \leq k \leq Q)$, where v is an arbitrary vector of $\mathbb{R}^d$, and $b_k^{(2)} = b_k^{(1)} + c$, $(1 \leq k \leq Q)$, where c is an arbitrary real, is also an optimal solution of Problem 1. To ensure the unicity, we thus impose the following additional constraints:

$$\begin{cases} \sum_{k=1}^{Q} w_k = 0_d \\ \sum_{k=1}^{Q} b_k = 0 \end{cases}$$

Taking into account these constraints, the SVM specified compares directly with the multi-category SVMs developed so far. Indeed, the only difference between these SVMs lies in the expression of the objective function, as can be seen in Table 1.

Computing the gradient of the Lagrangian function of the SVM proposed by Bredensteiner and Bennett and setting it equal to the null vector, we get at the

Table 1 Specifications of the different multi-category SVMs published so far.

SVM	Objective function	Add. constraints
Vapnik [76]	$\sum_{k=1}^{Q} \|w_k\|^2$	-
Bredensteiner and Bennett [10]	$\sum_{k<l}^{Q} \|w_k - w_l\|^2 + \sum_{k=1}^{Q} \|w_k\|^2$	-
This work	$\sum_{k<l}^{Q} \|w_k - w_l\|^2$	$\sum_{k=1}^{Q} w_k = 0_d$

optimum $\sum_{k=1}^{Q} w_k = 0_d$. This equality is also satisfied by the other multi-category SVMs [76, 78]. As a consequence, we get at the optimum:

$$\sum_{k<l}^{Q} \|w_k - w_l\|^2 = Q \sum_{k=1}^{Q} \|w_k\|^2$$

from which it springs that all the objective functions appearing in Table 1 are identical (modulo a multiplicative constant) and that the main multi-category SVMs exposed in literature so far are utterly equivalent to the one we have just specified.

To sum up, starting from a uniform strong law of large numbers, we have been able to derive in the framework of statistical learning theory the specifications of the sole multi-class SVM published so far, which had been "discovered" independently by several people. This result is interesting in its own right, since this rigorous justification was lacking, the reasons used by the aforementioned authors to support their choice being mainly the analogy with the bi-class case [76, 78, 10], and considerations regarding the regularization theory [10]. In the following section, we establish that our framework can be used to specify other models.

3.3 Different models obtained by changing the metric

All the computations performed up to now have been based on the use of the $d_{l_\infty, l_1(s_{2N})}$ pseudo-distance and $\|.\|_{l_\infty, l_1}$ norm. This is not a compulsory choice. Indeed, the bound of Theorem 1 can be modified very easily to take into account a change of metric. Furthermore, in order to optimize performance, selecting a specific (pseudo)-metric should result from a study of the nature of the problem at hand. This is precisely this degree of freedom which generates the family of models we have been dealing with. The primary limitation, to extend nicely Vapnik's bi-class SVM, is that the resulting training procedure must still amount to solving a convex programming problem. An exhaustive study of the different possibilities goes beyond the scope of this paper. In this section, we focus on a specific alternative, which will be used in the experiments detailed in the following sections.

Definition 6 *Let $\mathcal{F}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For a set s of points in $\mathcal{X}$, define the pseudo-metric $d_{l_\infty, l_\infty(s)}$ on $\mathcal{F}$ as:*

$$\forall (f, \bar{f}) \in \mathcal{F}^2,\ d_{l_\infty, l_\infty(s)}(f, \bar{f}) = \max_{x \in s} \max_{1 \leq k \leq Q} \left| f_k(x) - \bar{f}_k(x) \right|$$

For this definition of the pseudo-distance, and the corresponding norm, the following result, established by A. Elisseeff, holds:

Theorem 5 *If* $\max_{k<l} \|w_k - w_l\|$ *is bounded, then the expression of* $\mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, 2N)$ *with respect to the norm* $\|.\|_{l_{\infty,\infty}}$ *satisfies:*

$$\ln\left(\mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, 2N)\right) = O\left(Qd\ln(\frac{1}{\gamma})\right) \tag{7}$$

Building upon this formula, we can then specify the new multi-class SVM, the parameters of which are still the solution of a convex programming problem.

Problem 2

$$\min_{h\in\mathcal{H}} t^2 + \frac{C}{N}\sum_{i=1}^{N}\sum_{k=1}^{Q}\xi_{ik}$$

$$s.t. \begin{cases} (w_{C(x_i)} - w_k)^T x_i + b_{C(x_i)} - b_k \geq 1 - \xi_{ik}, & (1 \leq i \leq N), (1 \leq k \neq C(x_i) \leq Q) \\ \xi_{ik} \geq 0, & (1 \leq i \leq N), (1 \leq k \neq C(x_i) \leq Q) \\ \|w_k - w_l\|^2 \leq t^2, \ (1 \leq k < l \leq Q) \end{cases}$$

The choice between this model and the former one can for instance be based on the knowledge available regarding the domain in which the data live.

4 Implementation of the SVMs to Combine Protein Secondary Structure Prediction Methods

4.1 Characterization of the problem

To estimate the generalization capabilities of our SVMs, we used them to combine protein secondary structure prediction methods. Prediction of protein 3D structure from the primary sequence of amino acids is a very challenging task, for which no satisfactory solution is currently available. A step on this way is to predict the local conformation of the polypeptide chain, which is called the secondary structure. Protein secondary structure prediction is usually treated as a 3-class discrimination task, which consists in assigning a conformational state α-helix, β-strand or aperiodic (coil), to each residue (amino acid) of a sequence. Apart from the fact that this problem is of central importance in structural biology, it presents characteristics which make it highly attractive from the point of view of pattern recognition. First, large databases of protein chains are available, which makes it possible to assess the models developed to process them in a wide spectrum, ranging from small samples to asymptotic behaviour. Second, many different methods have been proposed to predict the secondary structure, as was already pointed out in Section 1. Third, combining these methods is not an easy task, since the risk of decreasing the training error while increasing the test error has been stressed by many specialists of the field. For all these reasons, we consider this problem to be a touchstone to assess ensemble methods.

4.2 Experimental protocol

We have implemented the SVMs associated with the $\|.\|_{l_\infty,l_1}$ norm (SVM1) and $\|.\|_{l_\infty,l_\infty}$ norm (SVM2) to combine the outputs of three of the most widely used secondary structure prediction methods: SOPMA [28], which uses multiple alignments, GOR IV [26], which is based on the formalism of the information theory and SIMPA96 [51], a nearest-neighbour method. To assess the resulting predictions,

we compared them with those of majority voting, a weighted average, optimal with respect to the least squares criterion, a Multi-Layer Perceptron (MLP) and the Multivariate Linear Regression Combiner (MLRC) introduced in [31, 34]. The MLR combiner requires the outputs of the experts to be class posterior probability estimates, and precisely to be non-negative and sum to unity. This is not the case with the prediction methods used here. In order to compare the combiners in a fair way, the outputs of the base classifiers were thus preliminary post-processed with the structure-to-structure filtering neural network described in [34]. To constitute the training and test sets, we selected a release of the PDBSELECT database [39] containing 629 chains. These chains are made up of 147518 residues. The secondary structure assignment was carried out according to DSSP [46]. In order to obtain unbiased estimates of the accuracy of the predictions, a variant of *stacked generalization* [80] was applied, to train in sequence the filtering networks and the combiners. The database was devided into seven disjoint parts of roughly equal size. Based on this splitting, a two-stage cross-validation procedure was implemented. Each subset was iteratively used as test set. The six remaining sets were then grouped by three in six different ways, in order to constitute as many pairs of disjoint training sets for the filtering networks and the combiners. In this variant, the initial leave-one-out cross-validation procedure was thus replaced with a more computationally efficient 6-fold cross-validation. This implementation of stacked generalization, although suboptimal, has been observed not to deteriorate the generalization performance, or more precisely the test error, which is consistent with other results, for instance those reported in [13]. The prediction accuracy was assessed by means of four standard measures: the percentage of correctly predicted residues Q_3 for a three-state description of secondary structure (helix, extended and aperiodic), Pearson's/Matthews' correlation coefficient C [52], the segment overlap measure Sov [64, 81] and the standard deviation in the secondary structure content σ. The Sov measure plays a specific role, since it evaluates the quality of the prediction with respect to the conformational segments, which is a criterion of primary importance for the task. The figures characterizing the behaviour of the individual methods, before and after filtering, have been gathered in Tables 2 and 3.

	GOR IV	SOPMA	SIMPA
Q_3	64.1	68.4	69.2
C_α	0.47	0.55	0.56
C_β	0.39	0.48	0.49
C_c	0.44	0.49	0.49
Sov	0.66	0.72	0.71
Sov_α	0.63	0.72	0.74
Sov_β	0.67	0.73	0.67
Sov_c	0.68	0.72	0.72
σ_α	13.9	10.8	10.8
σ_β	11.5	10.3	11.2
σ_c	9.4	9.9	11.6

Table 2 Initial relative prediction accuracy of individuals experts on a set of 629 non-homologous globular proteins from the PDBSELECT database.

4.3 Raw results of the combinations

Table 4 summarizes the relative performance of the different combiners. Figures given here correspond to SVMs with radial basis kernels ($\sigma = 0.1$) and $C = 10$.

	GOR IV	SOPMA	SIMPA
Q_3	66.5	69.7	69.4
C_α	0.51	0.58	0.57
C_β	0.43	0.49	0.49
C_c	0.46	0.50	0.49
Sov	0.68	0.71	0.70
Sov_α	0.67	0.73	0.72
Sov_β	0.64	0.68	0.66
Sov_c	0.70	0.72	0.71
σ_α	12.5	10.7	10.6
σ_β	11.6	11.1	10.7
σ_c	10.1	10.6	11.1

Table 3 Relative prediction accuracy of individuals experts on a set of 629 non-homologous globular proteins from the PDBSELECT database. Initial scores have been post-processed as was done in [34].

These values of the parameters were selected since they appeared to be "satisfactory" for both models. However, systematic experiments should be conducted in order to assess more precisely the influence of the parameterization. Furthermore, additional experiments performed with two different polynomial kernels seem to suggest that the choice of a particular kernel could have significant incidence on the prediction accuracy (data not shown). The training procedure consisted in a slight modification of the algorithm described in [23].

	vote	average	MLP	MLRC	SVM1	SVM2
Q_3	70.2	70.9	71.2	71.3	71.6	71.6
C_α	0.59	0.60	0.60	0.60	0.61	0.60
C_β	0.49	0.50	0.52	0.52	0.52	0.53
C_c	0.51	0.50	0.52	0.52	0.52	0.52
Sov	0.72	0.71	0.72	0.72	0.73	0.72
Sov_α	0.73	0.72	0.73	0.74	0.74	0.74
Sov_β	0.69	0.70	0.70	0.68	0.68	0.68
Sov_c	0.73	0.73	0.73	0.73	0.73	0.74
σ_α	10.5	10.0	10.1	10.3	10.6	10.6
σ_β	10.3	10.2	10.1	10.9	10.8	10.8
σ_c	10.1	10.3	10.5	11.4	11.2	11.1

Table 4 Relative prediction accuracy of combiners on a set of 629 non-homologous globular proteins from the PDBSELECT database.

The comparison of the predictive success of native methods and combinations illustrates the usefulness of the best combiners, which succeed in increasing significantly the recognition rate even though the spectrum of quality among the classifiers is wide. The SVMs obtain the best results, although the difference with the MLR combiner is slightly too low to be statistically significant (the corresponding confidence is only 0.93).

5 Post-Processiong of the Conformational Scores

Promising as they may seem, these results are not sufficient to determine to what extent the conformational scores computed by the SVMs can be useful for the biologist. A property one usually expects from these scores is the possibility to use them in higher-level treatments, or simply to provide a measure of reliability of the predictions, as was done in [63, 27, 61]. In order to evaluate the quality of the combiners with respect to these criteria, we post-processed their outputs with a dynamic programming algorithm inspired by [59], and first assessed for protein secondary prediction in [31]. To that end, the outputs of the SVMs had to be preliminary standardized. The outputs of the weighted average, the MLR combiner and the MLP were already class posterior probability estimates. The underlying Inhomogeneous Hidden Markov Model (IHMM) is depicted in Figure 1. It has

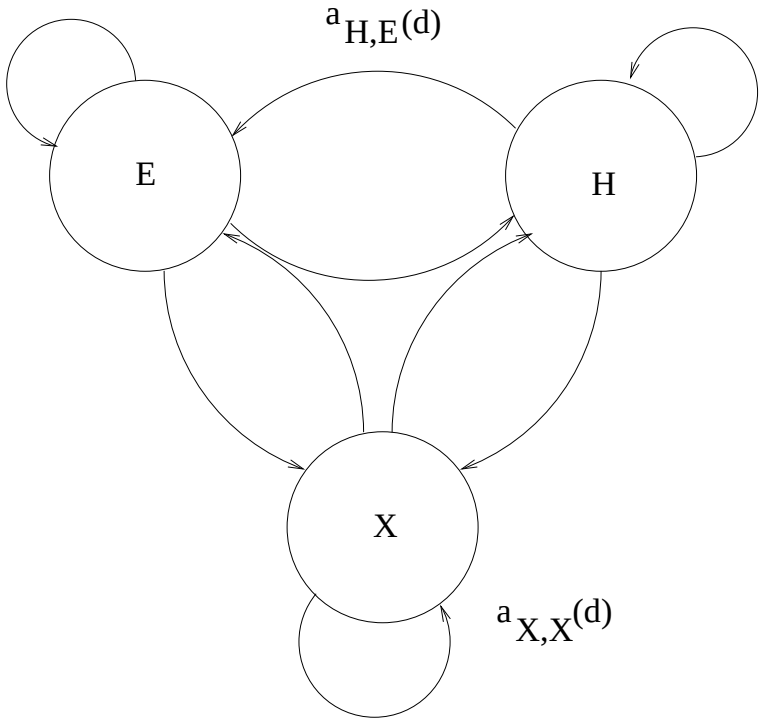

Figure 1 Topology of the IHMM used to post-process the outputs of the combiners.

three states, one for each conformational state. The observations are the residues of the primary structure. The specificity of the algorithm lies in the modeling of state durations. Instead of the standard stationary (first order) state transition probabilities, the terms $a_{ij}(d)$ are used, where the extra parameter d represents the duration spent in the current (conformational) state i. These probabilities are estimated by the corresponding frequencies observed on the training set, whereas the observation probability density functions are derived from the outputs of the combiners by means of Bayes' theorem. As can be seen in Table 5, the main advantage of such a post-processing is to narrow the gap between the length distributions of observed and predicted structural segments. However, its use also induces an improvement of the other standard measures of prediction accuracy. Indeed, the overall increase in recognition rate compared to the best results obtained so far on the same database, with the same experimental procedure [34], is now statistically significant with confidence exceeding 0.99. Once more, the SVMs obtain the best

	average	MLP	MLRC	SVM1	SVM2
Q_3	71.1	71.4	71.4	71.8	71.7
C_α	0.60	0.61	0.61	0.62	0.60
C_β	0.50	0.52	0.52	0.53	0.51
C_c	0.52	0.52	0.52	0.53	0.52
Sov	0.72	0.74	0.74	0.74	0.73
Sov_α	0.72	0.74	0.74	0.75	0.74
Sov_β	0.68	0.70	0.70	0.70	0.69
Sov_c	0.72	0.73	0.75	0.75	0.75
σ_α	10.6	11.8	10.8	10.7	10.7
σ_β	10.4	10.6	11.1	10.9	10.9
σ_c	10.6	10.8	11.6	11.8	11.7

Table 5 Quality of the predictions when the outputs of the combiners have been post-processed by an inhomogeneous DP algorithm.

results among the combiners, by a larger margin this time, which is statistically significant with confidence exceeding 0.98 for SVM1. This means that the values of their outputs carry some valuable information, and can for instance be used to estimate the class posterior probabilities. This property could prove to be very useful to incorporate them in hierarchical models far more complex than the one described here, such as the hybrid systems used in speech processing, user modeling or handwriting recognition.

6 Conclusion and Prospects

We have introduced a new family of multi-class SVMs. Unlike the previous works in the field [76, 78, 10], this study grounds the specification of the models directly on a uniform strong law of large numbers, with the consequence that the training procedure corresponds to an explicit implementation of the SRM inductive principle. Precisely, the training procedure amounts to minimizing an expression of the guaranteed risk derived from a uniform convergence result specifically established for Q-category discriminant models. This bound is tighter than those derived so far for models with multiple outputs, which should make the implementation of the SRM principle better justified in the context of multi-category discriminant analysis. Moreover, an appealing feature of these SVMs is the fact that they exhibit the properties which represent the main advantages of the bi-class SVM. Indeed, our models appear as natural generalizations of Vapnik's one, since their definitions are compatible with the extension of some of the main theorems regarding the generalization capacities [33]. Two of them have been implemented to combine protein secondary structure prediction methods. These combinations appear to give better performance than those resulting from the implementation of standard ensemble methods, the gain becoming statistically significant when the outputs are post-processed with a DP algorithm. The recognition rate of the overall system highlights the benefits that one could expect from generalizing the use of SVMs in the discriminant models performing tasks in biocomputing. So far, only bi-class SVMs, or variants of them, had been implemented in biology, for protein homology detection [42, 40, 41] or to process gene expression data [14].

Since we started this work, new prediction methods with superior accuracy became available [44, 4, 57]. We have begun to assess the influence of their inclusion

in different combinations [36]. The rudimentary hierarchical approach represented by the combination of the base classifiers and the DP algorithm can be developed in various ways. Currently, we are studying the use of N-Best algorithms [70], in order to provide the practitioner with alternative predictions among which he will be able to make his own choice, based on his expertise. Furthermore, we are implementing a system with multiple sliding windows, inspired by what has been done in [49]. Our long-term goal is to incorporate in our prediction systems the symbolic knowledge currently available for the task. This is the subject of collaborations with biologists. Concomitantly, we intend to derive new theoretical results, and specifically study the asymptotical behaviour of the different multi-class SVMs developed so far. In this respect, we see another benefit bestowed upon us by the use of models the capacity of which can be estimated precisely. It must be borne in mind that with the rapid strides made in molecular biology, especially in the field of genome sequencing, huge quantities of data will soon become available to underly the main predictive tasks in bioinformatics. As a consequence, implementing cross-validation to estimate the generalization error will become prohibitive, particularly in the context of hierarchical systems such as ours [31], trained with variants of stacked generalization [80]. Bounds similar to those presented in this article, and specifically distribution-dependent bounds, should then represent an efficient alternative, which could make it possible to save both in terms of `cpu` time and training data. These bounds could naturally be adapted, to make a better use of the specificities of model combination. At last, *concentration inequalities* [9] could provide us with new tools to meet all these goals.

Acknowledgments

The author gratefully acknowledges the support of the ESPRIT funded Working Group N. 27150 "Neural Networks and Computational Learning Theory". Most of the theory grounding the SVMs described in this paper has been developed in collaboration with Prof. H. Paugam-Moisy and A. Elisseeff. Experimental results could also be checked thanks to André's software. The author would like to thank Prof. G. Deléage, Dr J. Garnier, Dr C. Geourjon, Dr J.-F. Gibrat and Dr. J.-M. Levin for the availability of the software and predictions of the SOPMA, GOR IV and SIMPA96 methods. Thanks are also due to K. Bennett, J. Weston and C. Watkins for interesting discussions on multi-class SVMs and to A. Brun for carefully reading this manuscript.

References

[1] M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[2] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44:615–631, 1997.

[3] M. Anthony. Probabilistic analysis of learning in artificial neural networks: the PAC model and its variants. *Neural Computing Surveys*, 1:1–47, 1997.

[4] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15(11):937–946, 1999.

[5] P.L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536, 1998.

[6] J.M. Bates and C.W.J. Granger. The combination of forecasts. *Opl Res. Q.*, 20:451–468, 1969.

[7] V. Biou, J.-F. Gibrat, J.-M. Levin, B. Robson, and J. Garnier. Secondary structure prediction: combination of three different methods. *Prot. Eng.*, 2:185–191, 1988.

[8] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[9] S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. Technical Report NC2-TR-1999-057, NeuroCOLT2, 1999.

[10] E.J. Bredensteiner and K.P. Bennett. Multicategory Classification by Support Vector Machines. *Computational Optimization and Applications*, 12(1/3):53–79, 1999.

[11] L. Breiman. Bagging Predictors. *Machine Learning*, 24:123–140, 1996.

[12] L. Breiman. Bias, variance, and arcing classifiers. Technical Report 460, Statistics Department, University of California, Berkeley, 1996.

[13] L. Breiman. Stacked Regressions. *Machine Learning*, 24:49–64, 1996.

[14] M. Brown, W. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data using support vector machines. Technical report, University of California, Santa Cruz, 1999. (submitted for publication).

[15] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, June 1998.

[16] B. Carl and I. Stephani. *Entropy, compactness, and the approximation of operators.* Cambridge University Press, Cambridge, UK, 1990.

[17] R.T. Clemen and R.L. Winkler. Limits for the Precision and Value of Information from Dependent Sources. *Operations Research*, 33(2):427–442, 1985.

[18] C. Cortes and V.N. Vapnik. Support-Vector Networks. *Machine Learning*, 20:273–297, 1995.

[19] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, 2000.

[20] J.P. Dickinson. Some statistical results in the combination of forecasts. *Opl Res. Q.*, 24:253–260, 1973.

[21] J.P. Dickinson. Some comments on the combination of forecasts. *Opl Res. Q.*, 26(205–210), 1975.

[22] F. Eisenhaber, B. Persson, and P. Argos. Protein structure prediction: recognition of primary, secondary and tertiary structural features from amino acid sequence. *Crit. Rev. Biochem. Mol. Biol.*, 30:1–94, 1995.

[23] A. Elisseeff. *Etude de la complexité et contrôle de la capacité des systèmes d'apprentissage : SVM multi-classe, réseaux de régularisation et réseaux de neurones multicouches.* PhD thesis, ENS Lyon, 2000.

[24] A. Elisseeff, Y. Guermeur, and H. Paugam-Moisy. Margin error and generalization capabilities of multi-class discriminant models. Technical Report NC-TR-99-051, NeuroCOLT2, `http://www.neurocolt.com/abs/1999/abs99051.html`, 1999.

[25] R. Fletcher. *Practical Methods of Optimization*. Wiley, 1987.

[26] J. Garnier, J.-F. Gibrat, and B. Robson. GOR Method for Predicting Protein Secondary Structure from Amino Acid Sequence. *Methods Enzymol.*, 266:540–553, 1996.

[27] C. Geourjon and G. Deléage. SOPM: a self-optimized method for protein secondary structure prediction. *Protein Engineering*, 7(2):157–164, 1994.

[28] C. Geourjon and G. Deléage. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *CABIOS*, 11(6):681–684, 1995.

[29] F. Girosi, M.J. Jones, and T. Poggio. Priors, Stabilizers and Basis Functions: from regularization to radial, tensor and additive splines. Technical Report A.I. Memo N. 1430, C.B.C.L. Paper N. 75, MIT - AI laboratory, 1993.

[30] C.W.J. Granger. Combining Forecasts - Twenty Years Later. *Journal of Forecasting*, 8:167–173, 1989.

[31] Y. Guermeur. *Combinaison de classifieurs statistiques, application à la prédiction de la structure secondaire des protéines*. PhD thesis, Université Paris 6, 1997. (in French).

[32] Y. Guermeur, A. Elisseeff, and H. Paugam-Moisy. Estimating the sample complexity of a multi-class discriminant model. In *ICANN'99*, pages 310–315. IEE, 1999.

[33] Y. Guermeur, A. Elisseeff, and H. Paugam-Moisy. A new multi-class SVM based on a uniform convergence result. In *IJCNN'00*, volume IV, pages 183–188, 2000.

[34] Y. Guermeur, C. Geourjon, P. Gallinari, and G. Deléage. Improved performance in protein secondary structure prediction by inhomogeneous score combination. *Bioinformatics*, 15(5):413–421, 1999.

[35] Y. Guermeur and H. Paugam-Moisy. *Théorie de l'apprentissage de Vapnik et SVM, Support Vector Machines*, volume Apprentissage Automatique, chapter 4, pages 109–138. M. Sebban and G, Venturini, Hermès edition, 1999. (in French).

[36] Y. Guermeur and D. Zelus. Combining protein secondary structure prediction models with ensemble methods of optimal complexity. In *JOBIM'01*, 2001. (submitted).

[37] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

[38] D. Haussler and P.M. Long. A Generalization of Sauer's Lemma. *Journal of Combinatorial Theory, Series A*, 71:219–240, 1995.

[39] U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Sci.*, 3:522–524, 1994.

[40] T. Jaakola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 1999. (to appear).

[41] T. Jaakola, M. Diekhans, and D. Haussler. Using the Fisher kernel method to detect remote protein homologies. In *ISMB'99*, pages 149–158, 1999.

[42] T.S. Jaakola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, 1998.

[43] R.A. Jacobs. Methods for Combining Experts' Probability Assessments. *Neural Computation*, 7:867–888, 1995.

[44] D.T. Jones. Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. *J. Mol. Biol.*, 292:195–202, 1999.

[45] M.I. Jordan and R.A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.

[46] W. Kabsch and C. Sander. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*, 22(12):2577–2637, 1983.

[47] M.J. Kearns and R.E. Schapire. Efficient distribution-free learning of probabilistic concepts. In *Proceedings of the 31st Symposium on the Foundations of Computer Science*, pages 382–391, Los Alimatos, CA, 1990. IEEE Computer Society Press.

[48] A.N. Kolmogorov and V.M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in function spaces. *Amer. Math. Soc. Translations (2)*, 17:277–364, 1961.

[49] A. Krogh and S. Riis. Prediction of beta sheets in proteins. In *NIPS 8*, pages 917–923, 1996.

[50] M. LeBlanc and R. Tibshirani. Combining estimates in regression and classification. Technical Report 9318, Department of Preventive Medicine and Biostatistics and Department of Statistics, Univeristy of Toronto, Toronto, 1993.

[51] J.-M. Levin. Exploring the limits of nearest neighbour secondary structure prediction. *Protein Eng.*, 10(7):771–776, 1997.

[52] B.W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, 405:442–451, 1975.

[53] E. Mayoraz and E. Alpaydin. Support Vector Machines for Multi-Class Classification. Technical Report 98-06, IDIAP, 1998.

[54] B.K. Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989.

[55] F. Peng, R.A. Jacobs, and M.A. Tanner. Bayesian Inference in Mixture-of-Experts and Hierarchical Mixture-of-Experts Architectures. Technical report, Department of Biostatistics, University of Rochester, June 1994.

[56] M.P. Perrone. *Improving Regression Estimation: Averaging Methods for Variance Reduction with Extensions to General Convex Measure Optimisation.* PhD thesis, Department of Physics at Brown Univeristy, 1993.

[57] T.N. Petersen, C. Lundegaard, M. Nielsen, H. Bohr, J. Bohr, S. Brunak, G.P. Gippert, and O. Lund. Prediction of Protein Secondary Structure at 80% Accuracy. *PROTEINS: Structure, Function, and Genetics*, 41(1):17–20, 2000.

[58] D. Pollard. *Convergence of stochastic processes.* Springer-Verlag, N.Y., 1984.

[59] P. Ramesh and J.G. Wilpon. Modeling State Durations in Hidden Markov Models for Automatic Speech Recognition. In *ICASSP-92*, volume I, pages 381–384, 1992.

[60] M.D. Richard and R.P. Lippmann. Neural network classifiers estimate bayesian a posteriori probabilities. *Neural Computation*, 3:461–483, 1991.

[61] S. Riis and A. Krogh. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J. Comput. Biol.*, 3:163–183, 1996.

[62] B. Rost and S. O'Donoghue. Sisyphus and prediction of protein structure. *CABIOS*, 13:345–356, 1997.

[63] B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, 232:584–599, 1993.

[64] B. Rost, C. Sander, and R. Schneider. Redefining the Goals of Protein Secondary Structure Prediction. *J. Mol. Biol.*, 235:13–26, 1994.

[65] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

[66] B. Schölkopf, C.Burges, and V. Vapnik. Extracting support data for a given task. In *ICKDDM'95*, pages 252–257, 1995.

[67] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural Risk Minimization over Data-Dependent Hierarchies. Technical Report NC-TR-96-053, NeuroCOLT, 1996.

[68] J. Shawe-Taylor and N. Cristianini. Robust Bounds on Generalization from the Margin Distribution. Technical Report NC2-TR-1998-029, NeuroCOLT2, 1998.

[69] A.J. Smola. *Learning with Kernels.* PhD thesis, Technische Universität Berlin, 1998.

[70] V. Steinbiss. Sentence Hypotheses Generation in a Continuous-Speech Recognition System. In *Eurospeech-89*, pages 051–054, 1989.

[71] K. Tumer and J. Ghosh. Theoretical foundations of linear and order statistics combiners for neural pattern classifiers. Technical Report 95-02-98, The Computer and Vision Research Center, University of Texas, Austin, 1995.

[72] K. Tumer and J. Ghosh. Error Correlation and Error Reduction in Ensemble Classifiers. *Connection Science*, 8(3 & 4):385–404, 1996.

[73] K. Tumer and J. Ghosh. Estimating the Bayes Error Rate Through Classifier Combining. In *ICPR'96*, volume II, pages 695–699, 1996.

[74] V.N. Vapnik. *Estimation of dependences based on empirical data.* Springer-Verlag, N.Y, 1982.

[75] V.N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, N.Y., 1995.

[76] V.N. Vapnik. *Statistical learning theory.* John Wiley & Sons, Inc., N.Y., 1998.

[77] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[78] J. Weston and C. Watkins. Multi-class Support Vector Machines. Technical Report CSD-TR-98-04, Royal Holloway, Univeristy of London, Department of Computer Science, 1998.

[79] R.C. Williamson, A.J. Smola, and B. Schölkopf. Generalization Performance of Regularization Networks and Support Vector Machines *via* Entropy Numbers of Compact Operators. *IEEE Trans. on Information Theory*, 1999. (to appear).

[80] D.H. Wolpert. Stacked Generalization. *Neural Networks*, 5:241–259, 1992.

[81] A. Zemla, Č. Venclovas, K. Fidelis, and B. Rost. A modified definition of Sov, a seqment-based measure for protein secondary structure prediction assessment. *Proteins: Structure, Function, Genetics*, 34:220–223, 1999.

[82] X. Zhang, J.P. Mesirov, and D.L. Waltz. Hybrid System for Protein Secondary Structure Prediction. *J. Mol. Biol.*, 225:1049–1063, 1992.